DEFECT INSPECTION FROM SCRATCH TO PRODUCTION

Dr. Andrew Liu

Sr. Solution Architect

NVIDIA

AGENDA

Image Segmentation

- Fully Convolution Neural Network

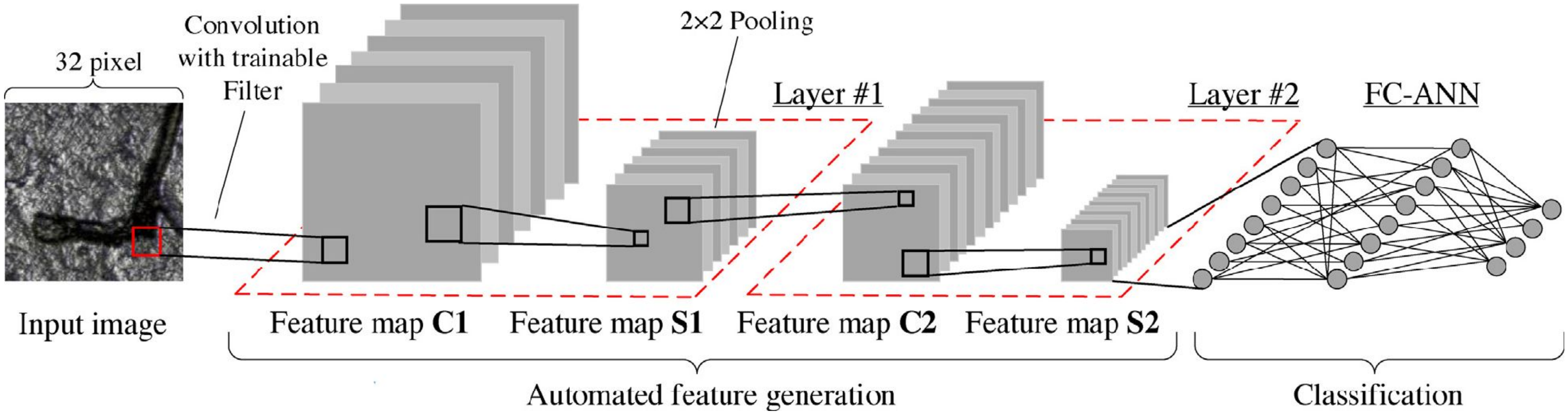
Defect Inspection

- Problem Define
- Data Preparation
- Deal with Imbalance data

Speed up with TensorRT

CNN STRUCTURE

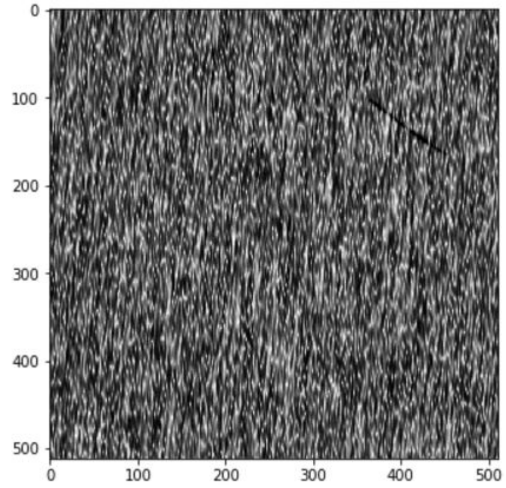
LeNet



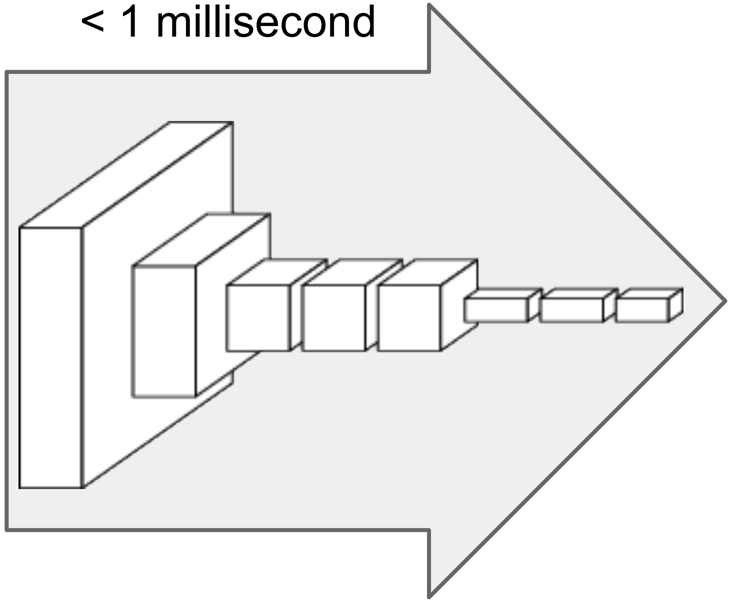
Source: Design of Deep Convolutional Neural Network Architectures for Automated Feature Extraction in Industrial Inspection, D. Weimer et al, 2016

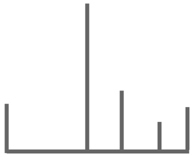
FULLY CONVOLUTION NEURAL NETWORK IMAGE SEGMENTATION

convnets perform classification



< 1 millisecond





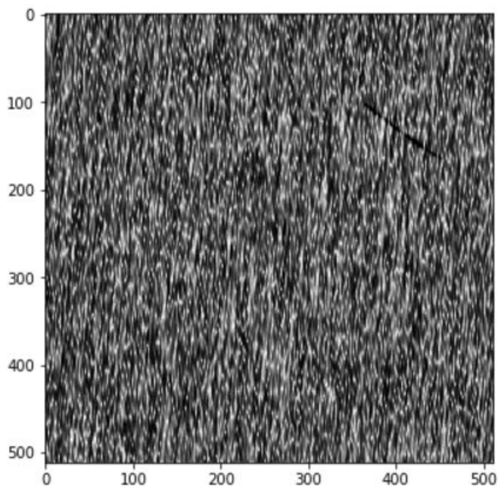
“defect class1”

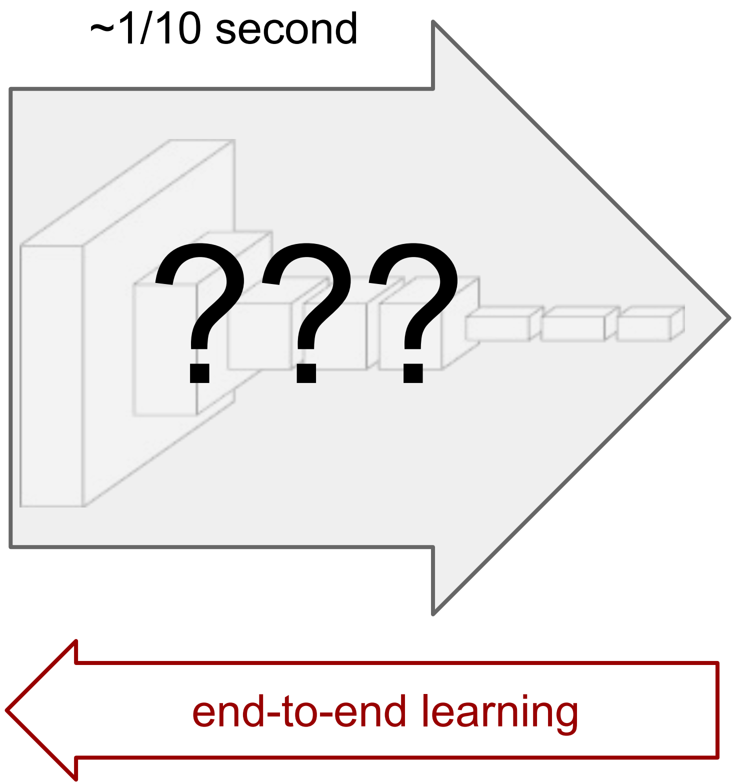
1000-dim vector

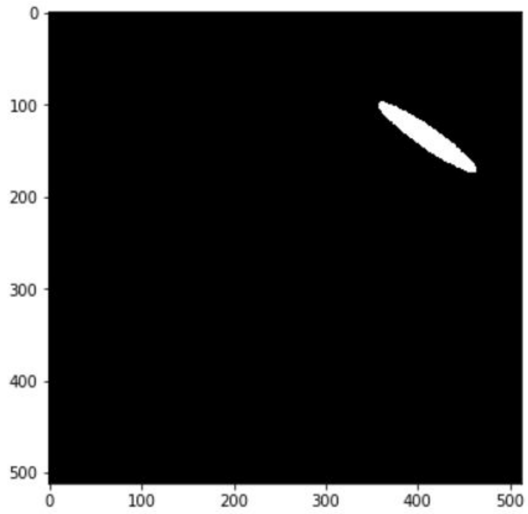
end-to-end learning

https://arxiv.org/abs/1605.06211

lots of pixels, little time?

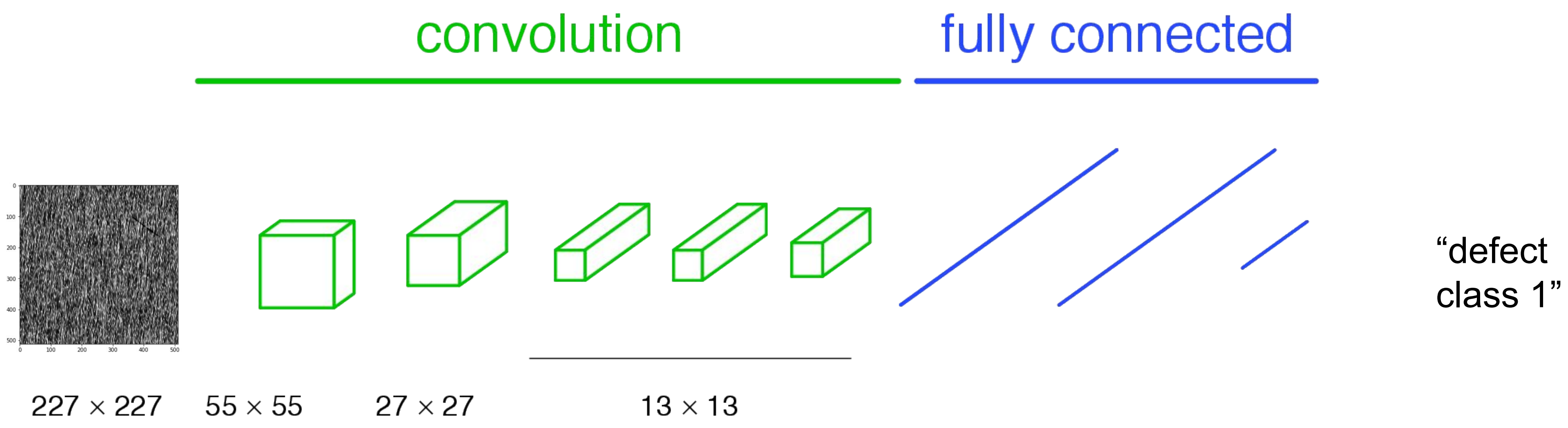






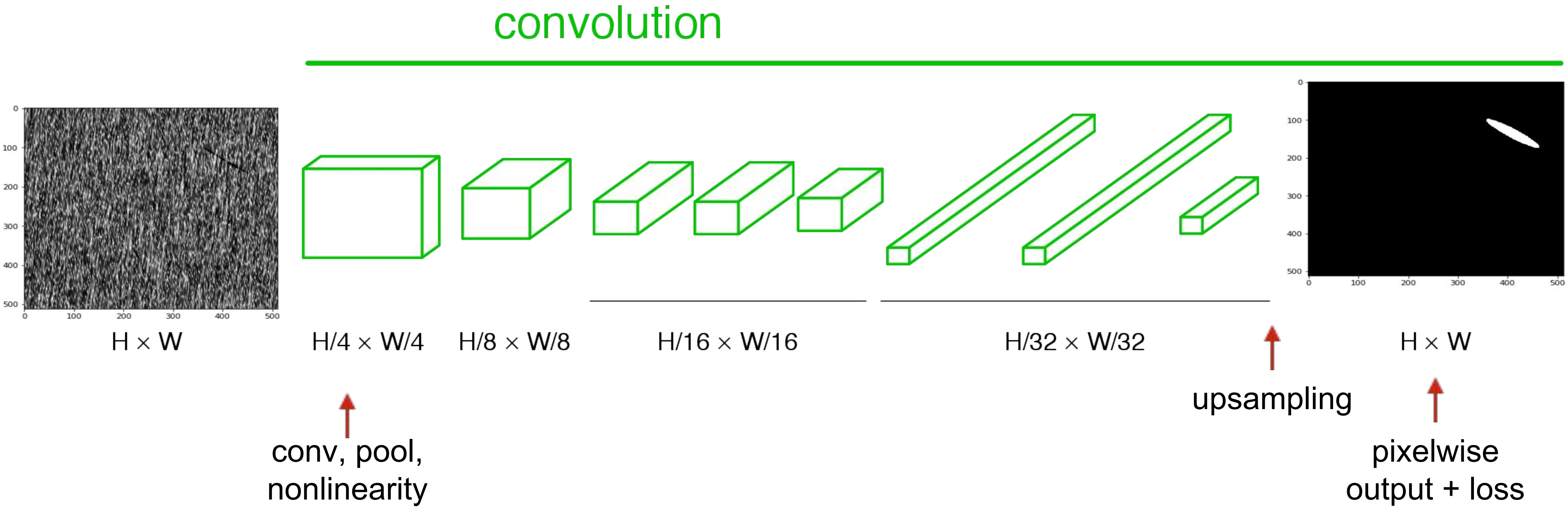
https://arxiv.org/abs/1605.06211

a classification network



https://arxiv.org/abs/1605.06211

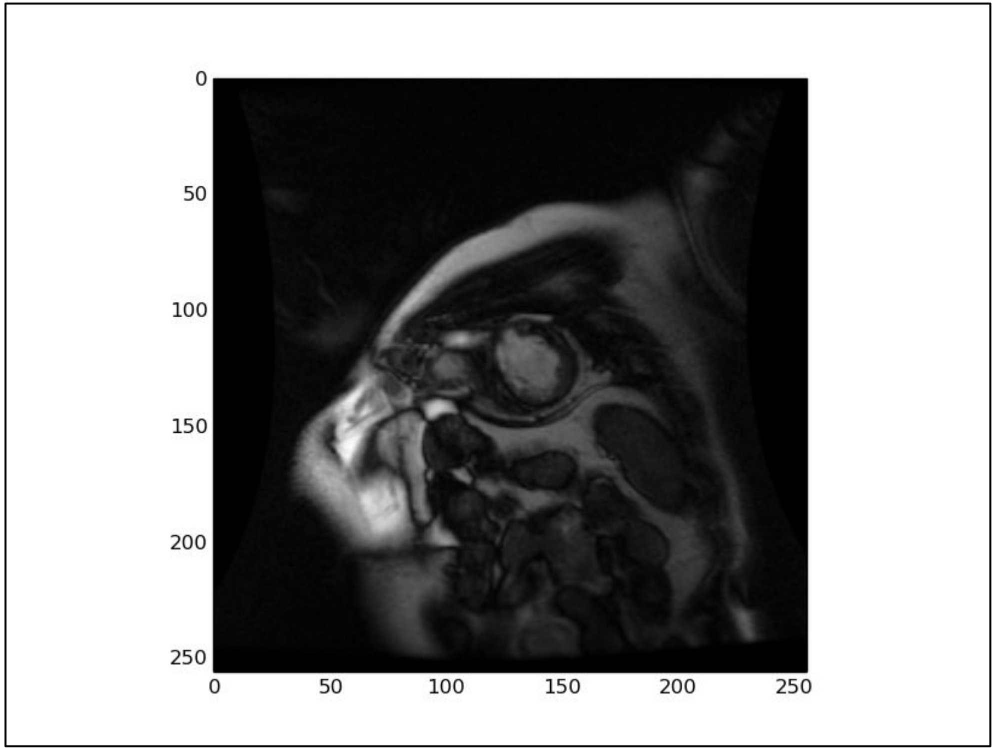
end-to-end, pixels-to-pixels network

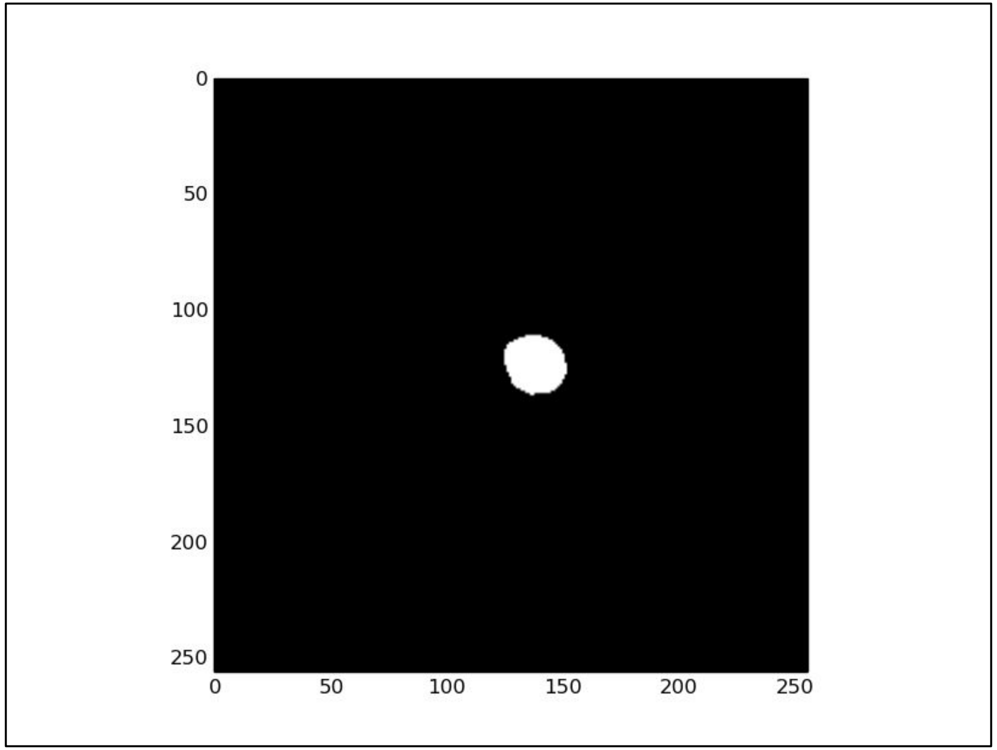


https://arxiv.org/abs/1605.06211

MRI image -> Left ventricle

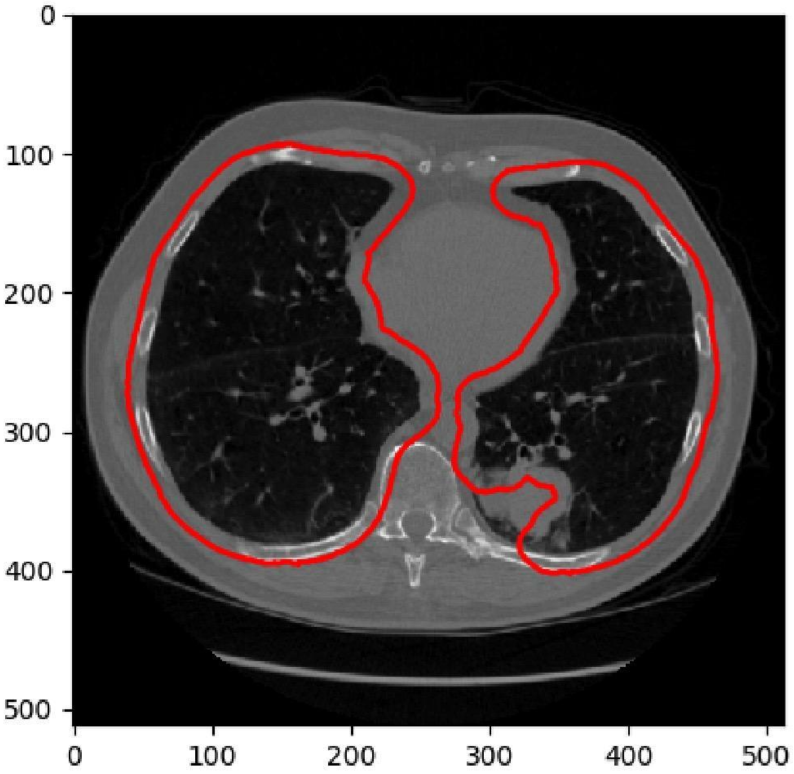
2nd Data Science BOWL competition

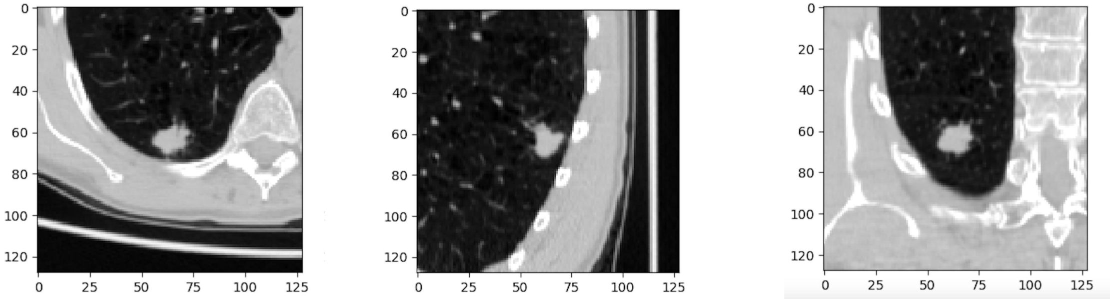




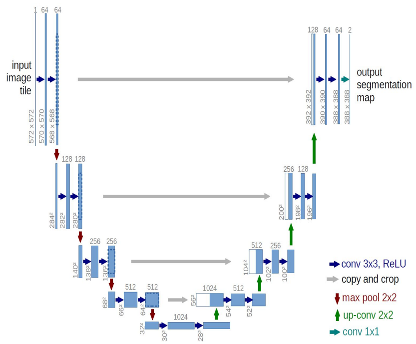
DATA SCIENCE BOWL 2017

Predicting Lung Cancer





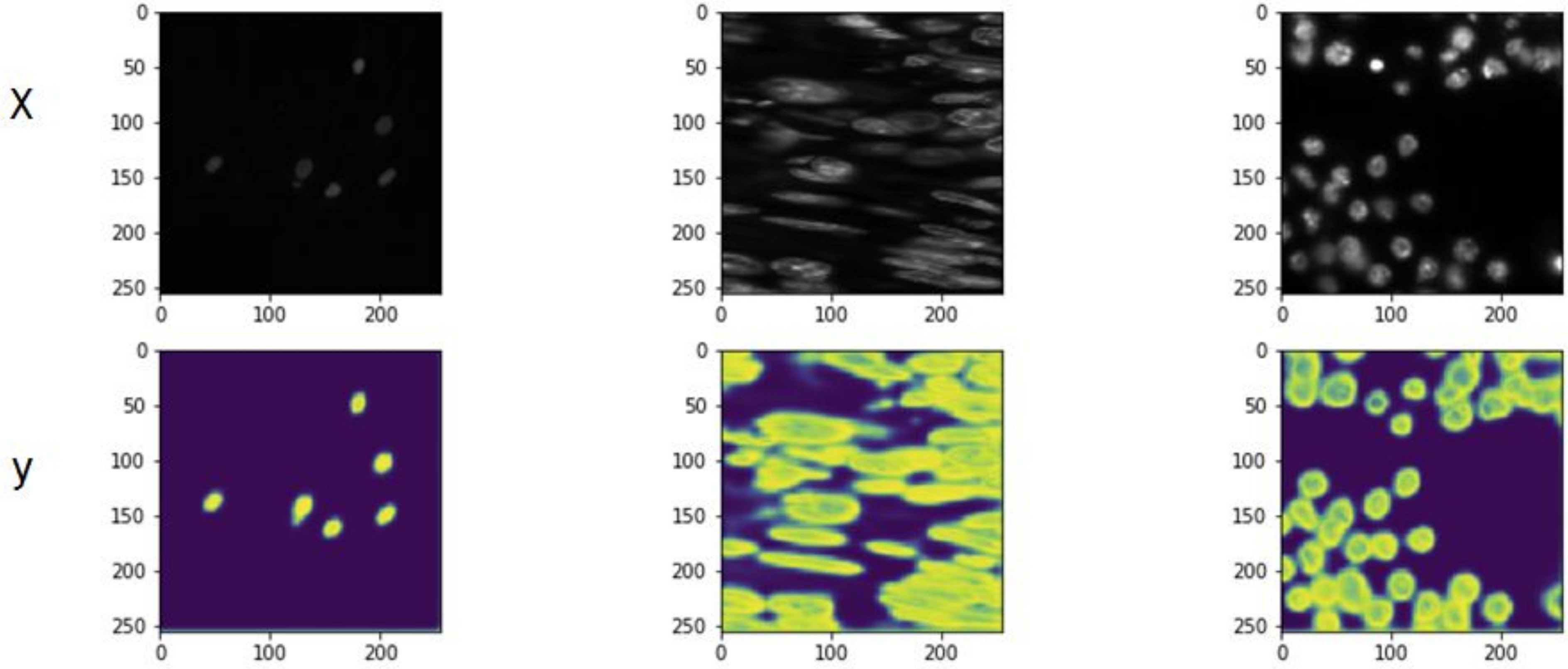
Lung segmentation



Nodule expert classification

DATA SCIENCE BOWL 2018

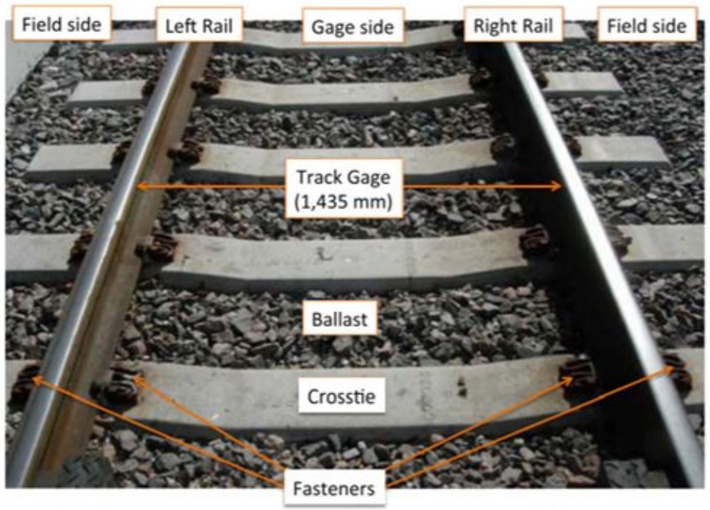
Predicting nuclei in divergent images

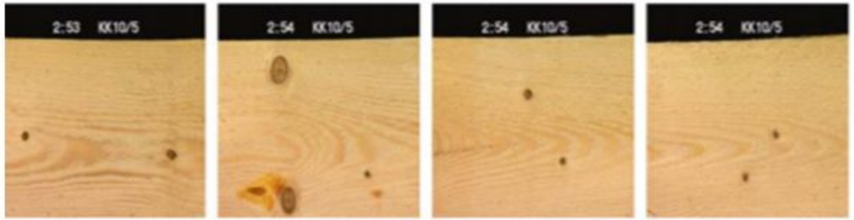


https://www.kaggle.com/c/data-science-bowl-2018/

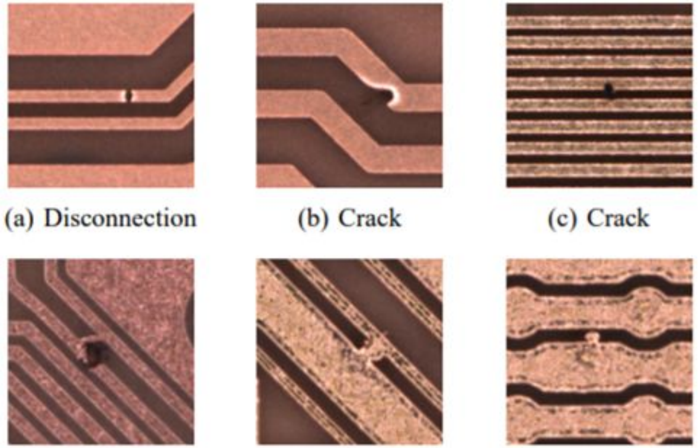
DEFECT INSPECTION

INDUSTRIAL DEFECT INSPECTION





(a) (b) (c) (d)



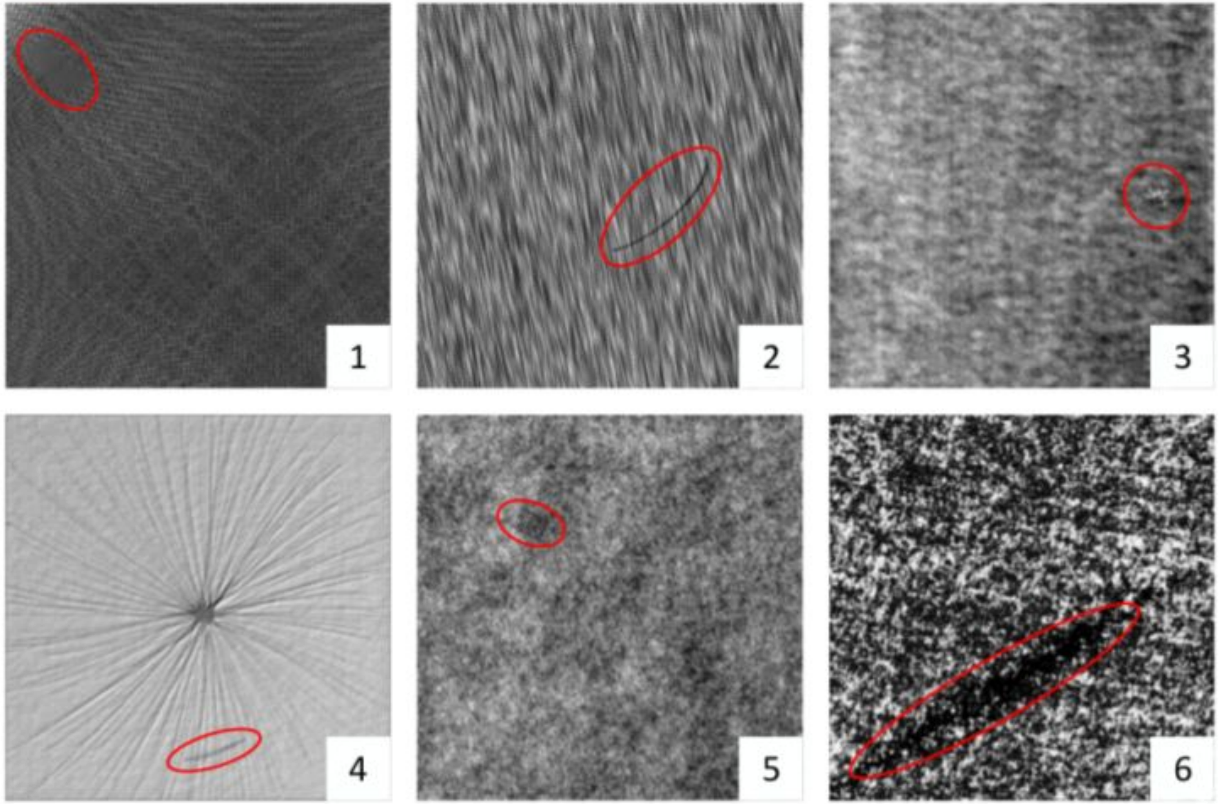
(a) Disconnection (b) Crack (c) Crack

(d) Connection (e) Connection (f) Projection

DATASET

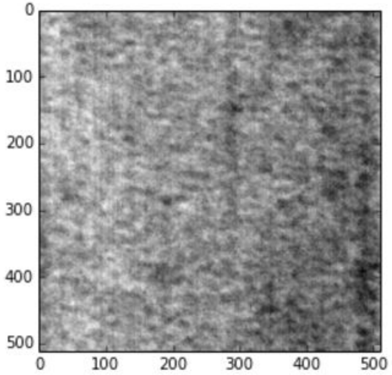
INDUSTRIAL OPTICAL INSPECTION

German Association for Pattern Recognition

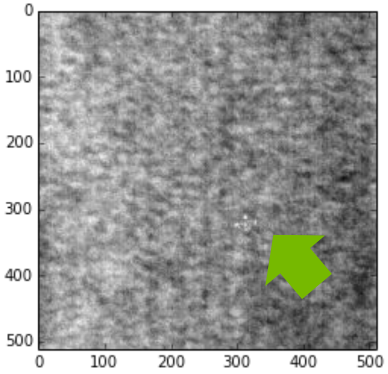


INDUSTRIAL OPTICAL INSPECTION

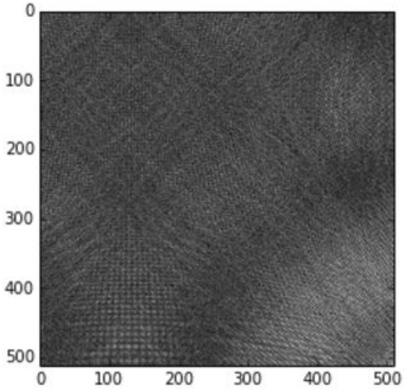
German Association for Pattern Recognition



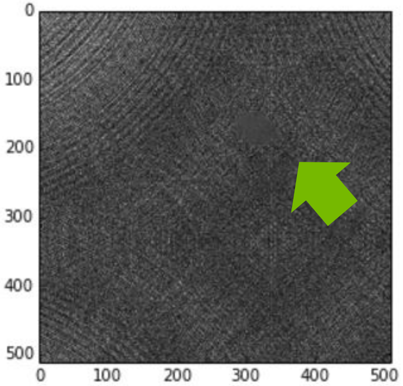
Pass



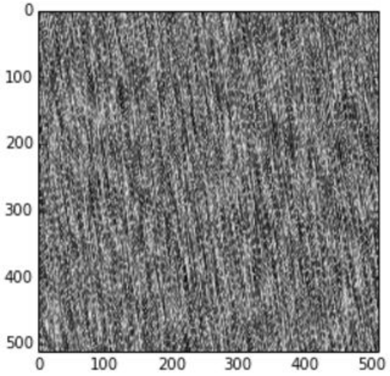
NG



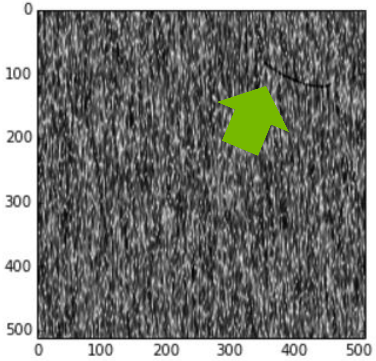
Pass



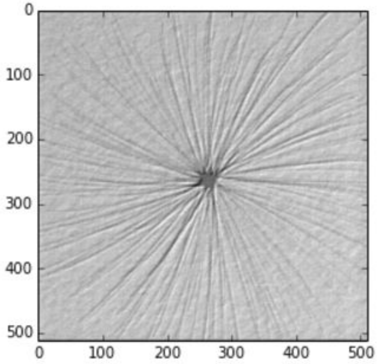
NG



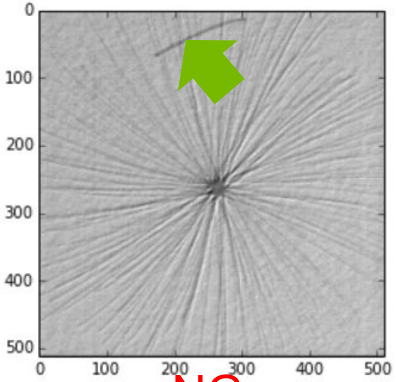
Pass



NG



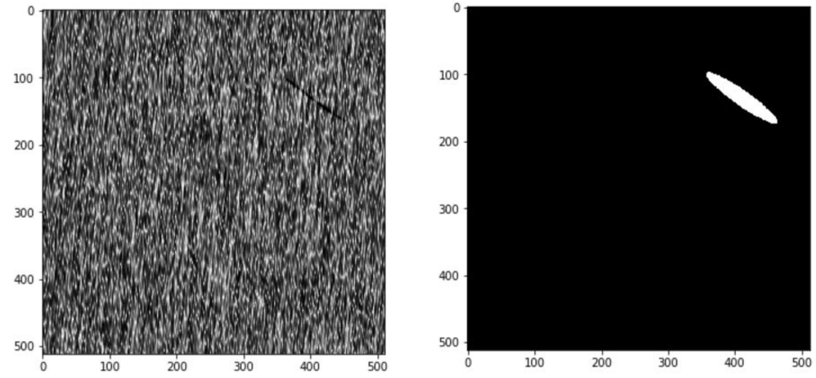
Pass



NG

DATA DETAILS

- Original images are 512 x 512 grayscale format
- Output is a tensor of size 512 x 512 x 1
    - Each pixel belongs to one of two classes
- Training set consist of 100 images
- Test set consist of 50 images

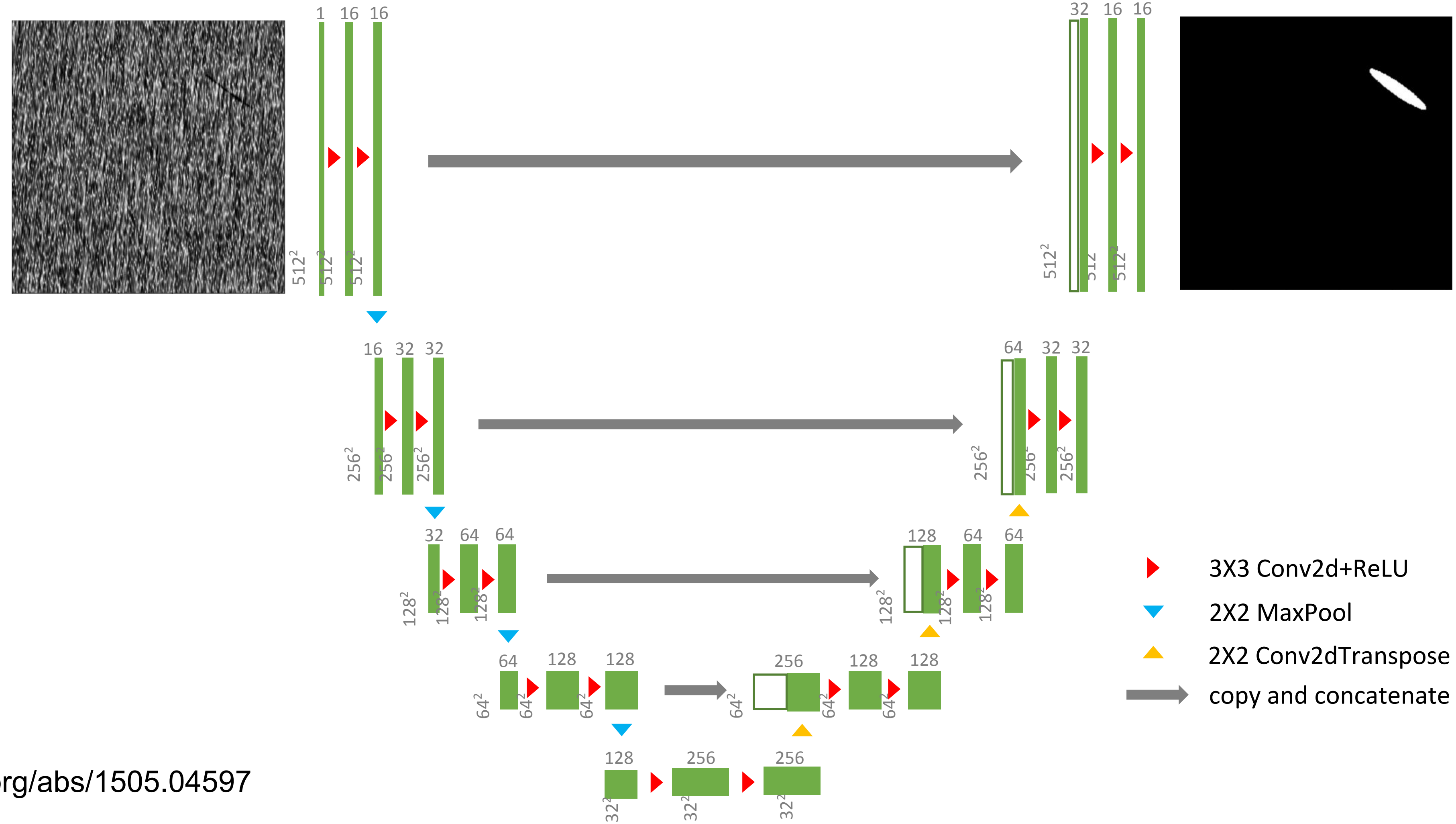


MODEL SET UP

Deconvolution layer

- Deconvolution (transpose convolution) layer
  - Up-sampling method to bring a smaller image data set back up to it's original size for final pixel classification
- Long et al (CVPR2015) has nice paper re: FCN for segmentation
  - Created FCNs from AlexNet and other canonical networks
- Zeiler et al (CVPR2010) describes deconvolution

U-Net structure



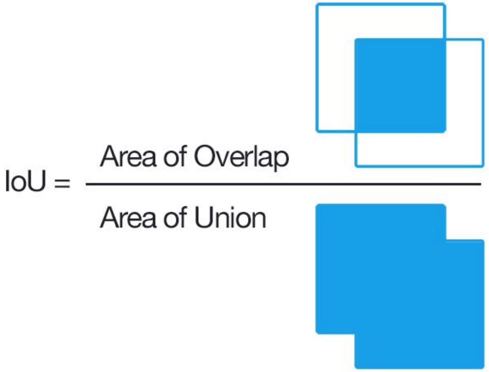
https://arxiv.org/abs/1505.04597

IMBALANCE DATA

Dice Metric

- Metric to compare the similarity of two samples:

$$\frac{2A_{nl}}{A_n + A_l}$$



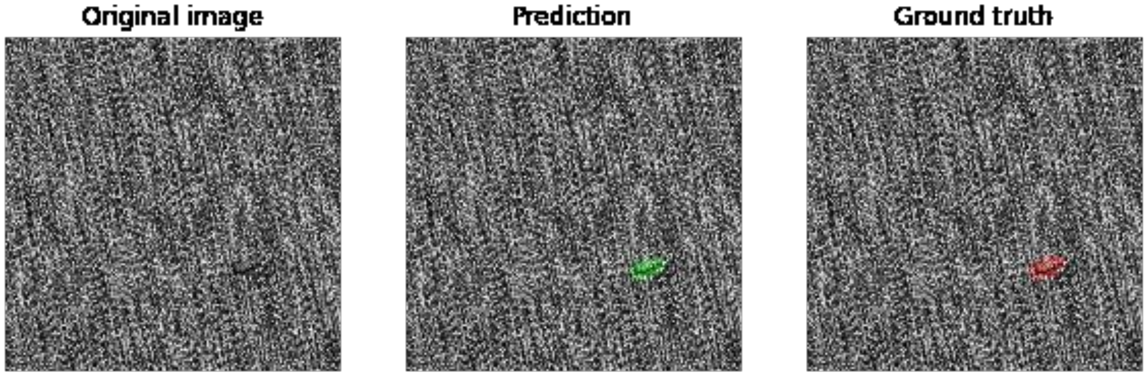
  - Where:
    - A_n is the area of the contour predicted by the network
    - A_l is the area of the contour from the label
    - A_{nl} is the intersection of the two
      - The area of the contour that is predicted correctly by the network
      - 1.0 means perfect score.

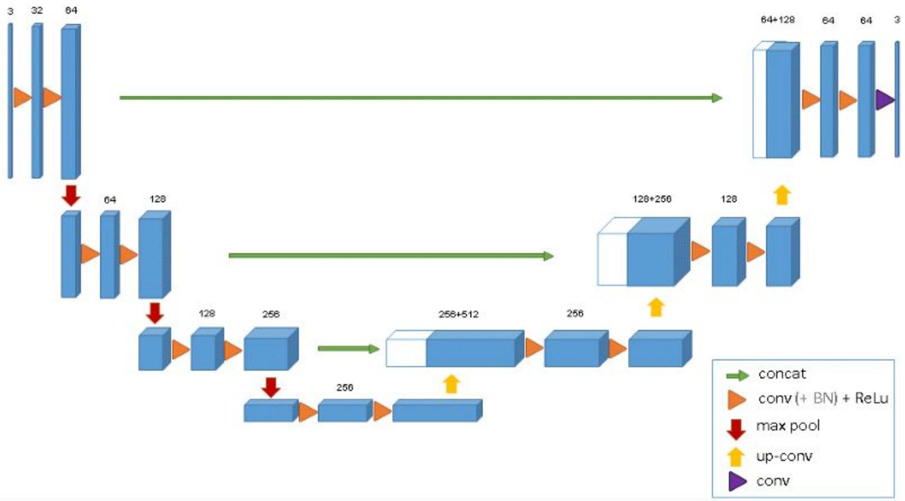
- More accurately compute how well we're predicting the contour against the label

- We can just count pixels to give us the respective areas

APPLICATION: INDUSTRIAL INSPECTION

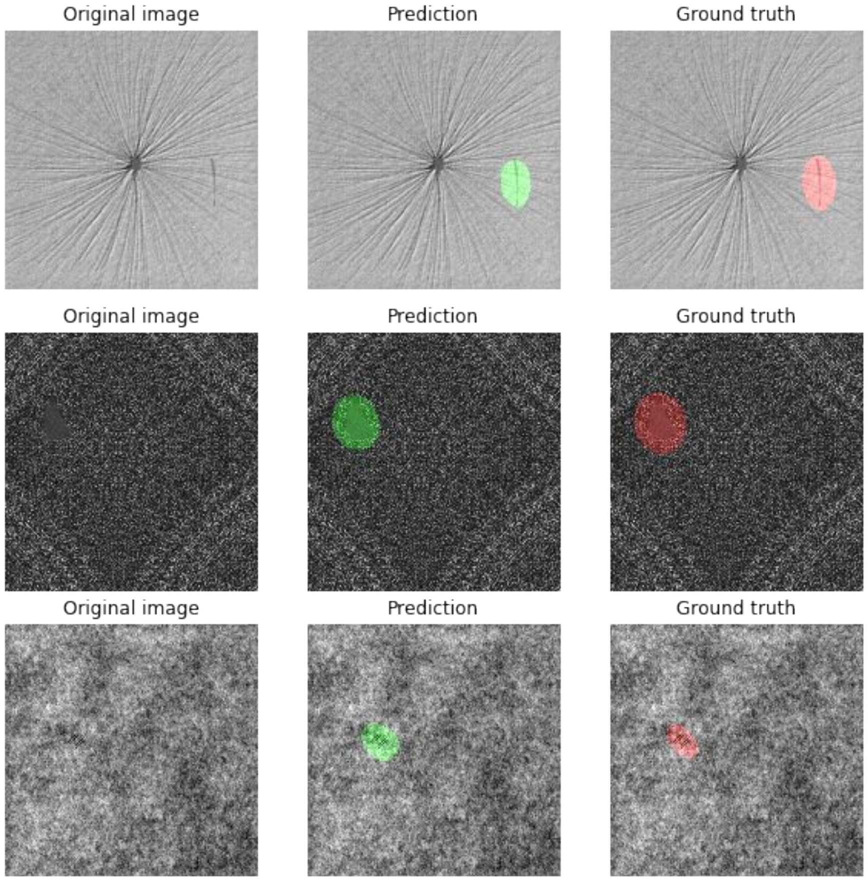
NVIDIA





FINAL DECISION

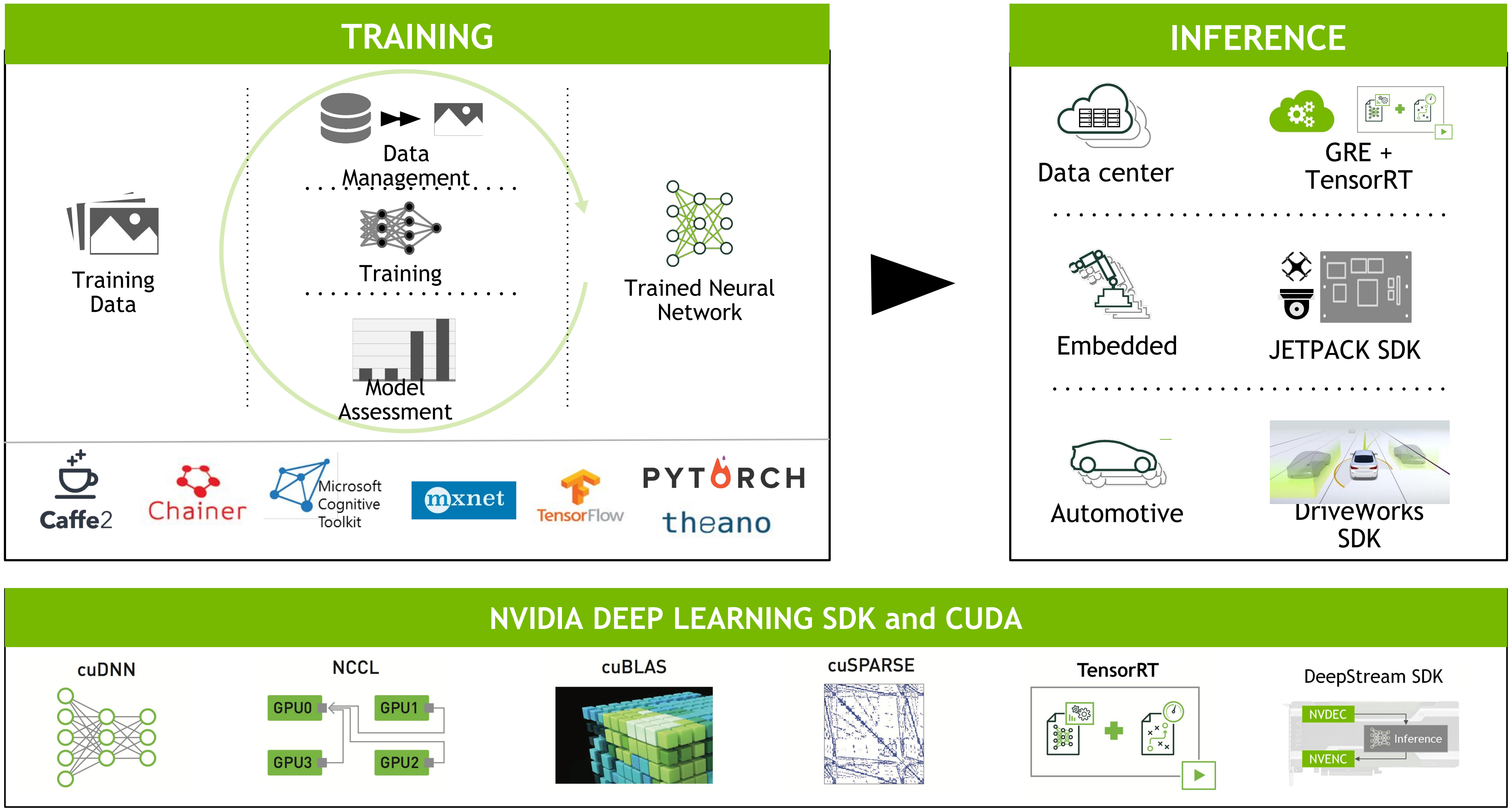
Plus Human Logic



Size, Position, … etc

PRODUCTION

NVIDIA DEEP LEARNING SOFTWARE PLATFORM



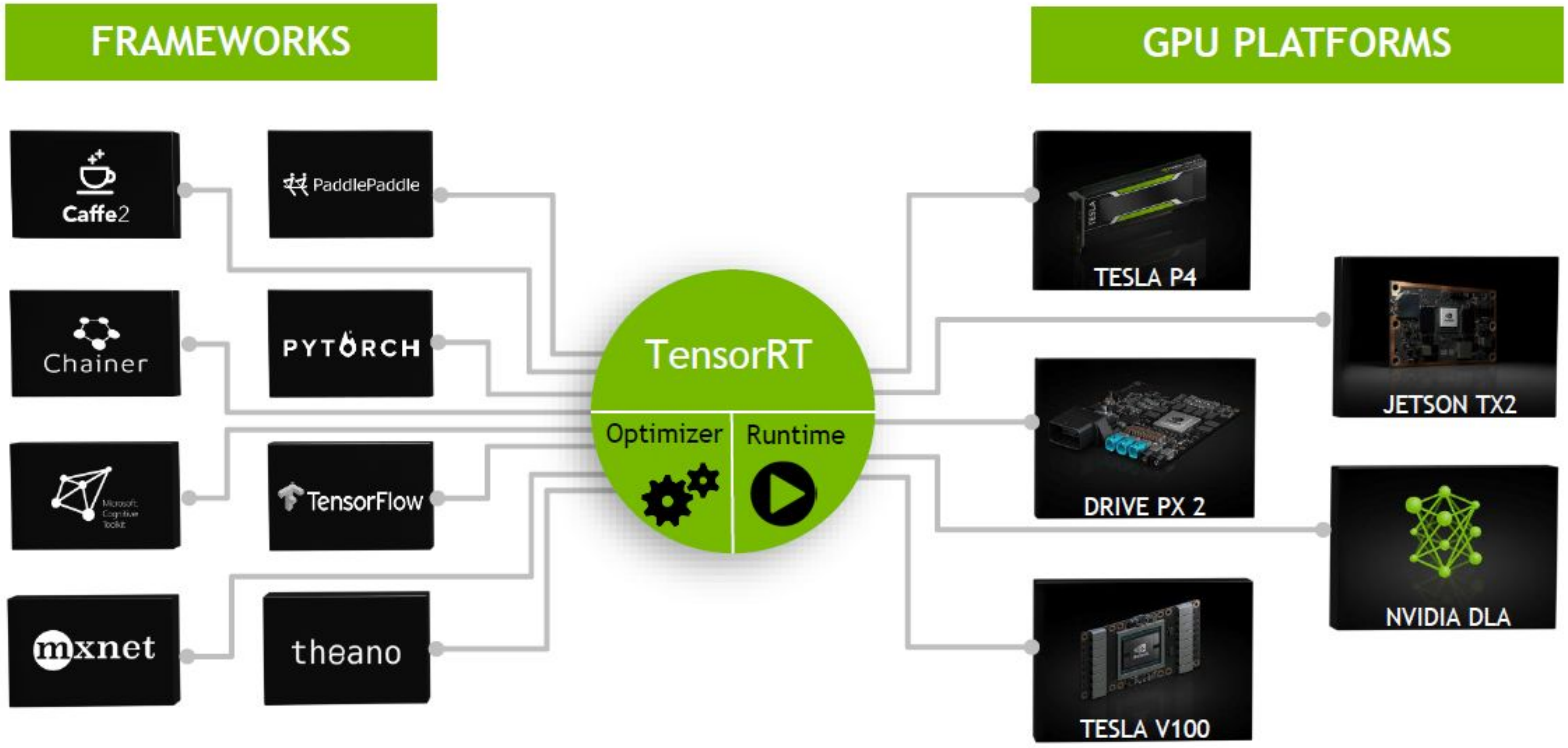
developer.nvidia.com/deep-learning-software

CHALLENGES DURING PRODUCTION

Requirement	Challenges
High Throughput	**Unable to processing high-volume, high-velocity data** ➢ Impact: Increased cost ($, time) per inference
Low Response Time	**Applications don't deliver real-time results** ➢ Impact: Negatively affects user experience (voice recognition, personalized recommendations, real-time object detection)
Power and Memory Efficiency	**Inefficient applications** ➢ Impact: Increased cost (running and cooling), makes deployment infeasible
Deployment-Grade Solution	**Research frameworks not designed for production** ➢ Impact: Framework overhead and dependencies increases time to solution and affects productivity

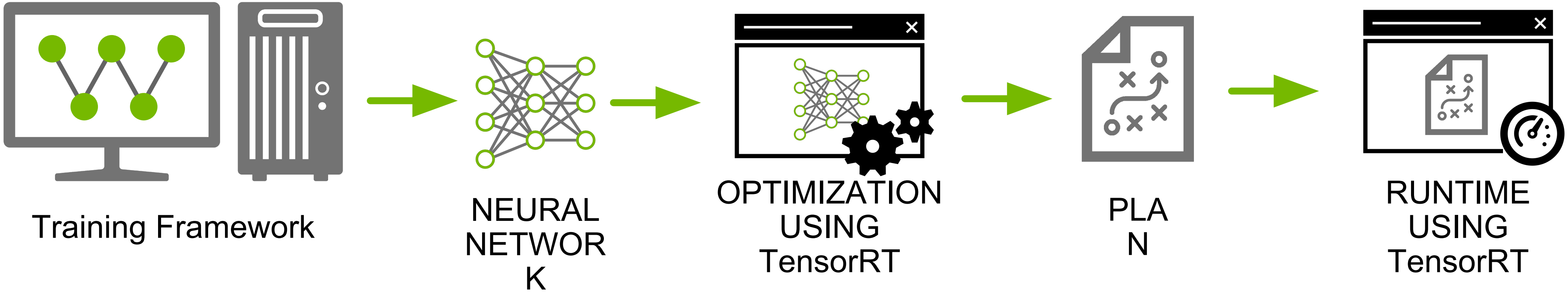
NVIDIA TENSORRT

Programmable Inference Accelerator



TENSORRT

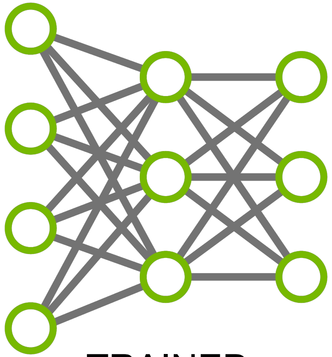
Workflow



developer.nvidia.com/tensorrt

TENSORRT

Optimizations



TRAINED
NEURAL
NETWORK

- **Fuse network layers**
- **Eliminate concatenation layers**
- **Kernel specialization**
- **Auto-tuning for target platform**
- **Tuned for given batch size**

OPTIMIZED INFERENCE RUNTIME

CHALLENGES ADDRESSED BY TENSORRT

Requirement	*TensorRT Delivers*
High Throughput **Low Response Time**	**Maximizes inference performance on NVIDIA GPUs** ➢ INT8, FP16 Precision Calibration, Layer & Tensor Fusion, Kernel Auto-Tuning ➢ Up to 40x Faster than CPU-Only inference and 18x faster inference of TensorFlow models ➢ Under 7ms real-time latency
Power and Memory Efficiency	**Performs target specific optimizations** ➢ Platform specific kernels for Embedded (Jetson), Datacenter (Tesla GPUs) and Automotive (DrivePX) ➢ Dynamic Tensor Memory management improves memory re-use
Deployment-Grade Solution	**Designed for production environments** ➢ No framework overhead, minimal dependencies ➢ Multiple frameworks, Network Definition API ➢ C++, Python API, Customer Layer API

THANKS!